An Analysis of Perplexity to Reveal the Effects of Alzheimer's Disease on Language

Sebastian Wankerl[1], *Elmar Nöth*[1], *Stefan Evert*[2]

[1] Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen
[2] Professur für Korpuslinguistik, Friedrich-Alexander-Universität Erlangen-Nürnberg, Bismarckstr. 6, 91054 Erlangen
Email: {sebastian.wankerl,elmar.noeth,stefan.evert}@fau.de
Web: [1]www5.cs.fau.de, [2]linguistik.fau.de

Abstract

The effects of Alzheimer's disease on language have been examined in the past years. However, most of the methods described so far rely on language specific tools like part-of-speech taggers or parsers. We describe a novel approach based on N-gram models which has the benefit of being able to be used for any language which has a writing system. We evaluated our approach on the works of novelist Iris Murdoch whose last novel was written in the stage of early Alzheimer's disease. We show that our approach reveals the influence of Alzheimer's disease similarly well to the approaches that have previously been described in literature.

1 Introduction

Alzheimer's disease is one of the diseases for which exists no cure to the present day. For a better understanding of Alzheimer's disease—which might eventually lead to an effective treatment—it is therefore worthwhile to detect and understand every part of human capabilities that is changed under the influence of the ailment.

The effects of Alzheimer's disease on the production of language have been studied thoroughly in the past years, considering speech[1], as well as creative writing[2–4]. For the latter goal, however, there is only little reliable data available.

1.1 Iris Murdoch

A widely used corpus for examining the influences of Alzheimer's disease on creative writing are the works of Iris Murdoch, a reputable British novelist. Born in 1919, she published her first novel "Under the Net" in 1954 and her last novel "Jackson's Dilemma" in 1995. All in all, she published a total of 26 novels as well as several non-fictional texts on philosophy. Her novels usually cover a wide range of topics and were, especially at the peak of her career, known for their moral controversies and philosophical depth[5].

Iris Murdoch was diagnosed to suffer from Alzheimer's disease in 1996 and died in 1999. The diagnosis was later confirmed by a post-mortem examination[2]. In contrary to her previous works, her last novel was received "without enthusiasm" by literary critic and Murdoch stated that she had experienced a writer's block during its creation[2]. Moreover, Murdoch was known for allowing no interference from editors[2]. Hence, it is likely that the language in this novel is highly influenced by symptoms of cognitive decline which were already present at this time.

1.2 Related Works

Garrard et al. [2] performed a study on Iris Murdoch's writing, examining three of her novels: the first, the last, and one published at the peak of her writing career in 1978, named "The Sea, The Sea". They evaluated various measures, including vocabulary growth, number of repetitions, the mean number of clauses per sentence, familiarity of used words, and length of used words. With the exception of vocabulary growth, these measures were, however, only evaluated on small samples of the whole text.

They found some noticeable differences between the last novel and the earlier ones. Among other things, it has a slower increasing vocabulary growth curve, more repetitions within small intervals and generally consists of words of higher familiarity. Because of the limitations described, it is however questionable whether the obtained results are reliable[4].

Pakhomov et al. [3] analyzed Murdoch's writing under mainly syntactic aspects. They evaluated various syntactic complexity measures, using the same novels as [2] as well as the next-to-last novel "The Green Knight". However, like [2] they only evaluated small passages. For some evaluated measures, they observed a successive decrease of the syntactic complexity throughout the novels.

Le et al. [4] analyzed 20 of Murdoch's novels, as well as novels of two other writers. They borrowed some of the measures from [2, 3], adding additional tests like frequency of passive voice and D-Level. Furthermore, they processed the full novels. Among others, they found a sharp decrease in Murdoch's vocabulary size, a tendency to prefer verb tokens over noun tokens, as well as a decline in the frequency of the passive voice. In contrary to [3], evaluating the full texts they did not find a decrease of the syntactic complexity in Murdoch's last novel and disproved the usefulness of syntactic complexity measures for the diagnosis of Alzheimer's.

1.3 Our Goal

We propose the construction of n-gram language models with subsequent evaluation of the perplexity for the detection of signs of Alzheimer's disease in language. This approach has the advantage of not requiring a specific language because it does not depend on any external data like word frequency or familiarity, respectively tools like part-of-speech taggers or syntax parsers.

2 Theoretical Background

2.1 N-Gram Language Models

N-gram language models, see [6, 7], are a frequently used technique in fields like speech recognition or machine translation. Given a sentence $S = (w_1, \ldots, w_N)$ of words in a test corpus, one wants to estimate its probability based on a training corpus. In the simplest case, one can count the

words in the training corpus and assign them an adequate probability. This is called a unigram and the probability of S equals the product of the probabilities of each of its words.

The drawback of a unigram model is that it doesn't comprise any contextual information. On the other hand, considering the whole history of each word leads to very unique probabilities and since it is unlikely that the same sequence of words reoccurs in the test data, this approach is not practicable.

Taking into account the Markov assumption one can limit the history of each word to a few predecessors. Such a model that contains a history of $n-1$ words is called an n-gram. For a sequence $S = (w_1, \ldots, w_k)$ of k words, the probability is then evaluated by

$$p(S) = \prod_{i=1}^{k} P(w_i | w_{i-n+1}, \ldots, w_{i-1}) \quad (1)$$

To calculate the probability of the first words of a sentence, an artificial token which marks the beginning of a sentence is created. Moreover, the end of a sentence is expressed by a special token, too. An n-gram that takes into account a history of length one is called a bigram, for history of length two it is called a trigram.

When training a language model, it is crucial to reshape the probabilities estimated from relative n-gram frequencies such that n-grams which were not observed in the training data receive a probability greater than zero. This is necessary since one cannot distinguish between n-grams that are impossible and those which did not appear in the training corpus due to its size. Various smoothing techniques are used in literature.

2.2 Perplexity

Perplexity is used to evaluate how well an n-gram model fits the test data. The lower the perplexity, the better the test data can be predicted by the model. For a sequence of words $S = (w_1, \ldots, w_k)$ of test data, the perplexity is calculated by

$$PPL(S) = P(S)^{-\frac{1}{k}} \quad (2)$$

The perplexity can also be seen as the weighted average branching factor of the data, that is the number of possible next words that may follow any word in the test data.

3 Material and Methods

3.1 Data

We processed 19 of Murdoch's 26 novels, including the first two and the nine last novels. These last novels cover a stretch of almost 20 years of Murdoch's writing career, ranging from 1976–1995.

In contrary to the studies described in 1.2, we extracted the text directly from the ebook in epub file format and thereby bypassed the usage and possible errors of an OCR software.

[2, 3] excluded direct discourse from their data since it may differ in style. Thus, to be able to recognize the respective passages, we adapted the punctuation to get distinct markers for start and end of direct discourse. The novel "A Severed Head" (published in 1961) where the creation of those markers was impossible due to the formatting of the book, has, consequently, been excluded from the studies when the detection of direct discourse was necessary.

For tokenization and sentence splitting, we used Stanford CoreNLP[8]. Additionally, we calculated the number of types as well as the number of tokens for each book.

3.2 Experiments

Of each novel, we randomly partitioned its sentences into ten bins and used nine of them as training data and the left one for testing.

Subsequently, for every novel we built two n-gram language models out of its training data, a bigram model and trigram model. For that purpose, we used the SRILM-toolkit[9]. For smoothing we used both Witten-Bell discounting[10] as well as Kneser-Ney smoothing[11]. We enabled interpolation for both smoothing methods. We set the words from the training set as the vocabulary and, consequently, treated those words that only occurred in the test set as out-of-vocabulary.

Afterwards, using [9] we calculate the perplexity of the training model using the sentences from the test set as input sequence. To ensure the validity of the results, we repeated the above process ten times (ten-fold cross-validation) and used the mean perplexity as result.

We did the analysis both for the complete texts as well as the texts excluding direct discourse, like [2, 3]. Beside the differences in style, we expect the language used in direct speech to be generally simpler than the language used for, e.g., descriptions of scenes or characters' thoughts. We therefore suspect the writer's talent to be expressed especially in the latter parts. As a result of this, the cognitive decline caused by Alzheimer's disease is likely to affect these passages especially.

In addition to the n-gram models, we also calculated the number of types normalized from 60,000 tokens as well as the ratio of noun and verb tokens. The first one resembles the vocabulary growth curve, the latter should increase under the influence of dementia due to the finding by [4].

4 Results

Figure 1 shows the perplexity values and contrasts them with the noun/verb ratio as well as the normalized type count of the novels, including direct discourse. For clarity, we show perplexity and each of the other measures separately. Considering the perplexity values, we omitted the results of Witten-Bell discounting as they were similar to Kneser-Ney. It is clearly visible that the perplexity values of the trigram model are generally lower than those of the corresponding bigram model. This corresponds to findings from literature [11].

For the last novels, there is a strong decreasing tendency which is unique throughout the whole writing career. This trend culminates in the last novel whose perplexity values are significantly lower, regardless of the order of the language model.

As pointed out in 2.2, a low perplexity indicates a better predictability. In our case, one can conclude that the language is less diverse throughout a book with a low perplexity compared to a book with a higher perplexity. That means that there are, for example, more reoccurring words and phrases between the various parts of the book. All in all, the particularly low perplexity of the last novel reveals an impoverished language that can be attributed to Alzheimer's disease.

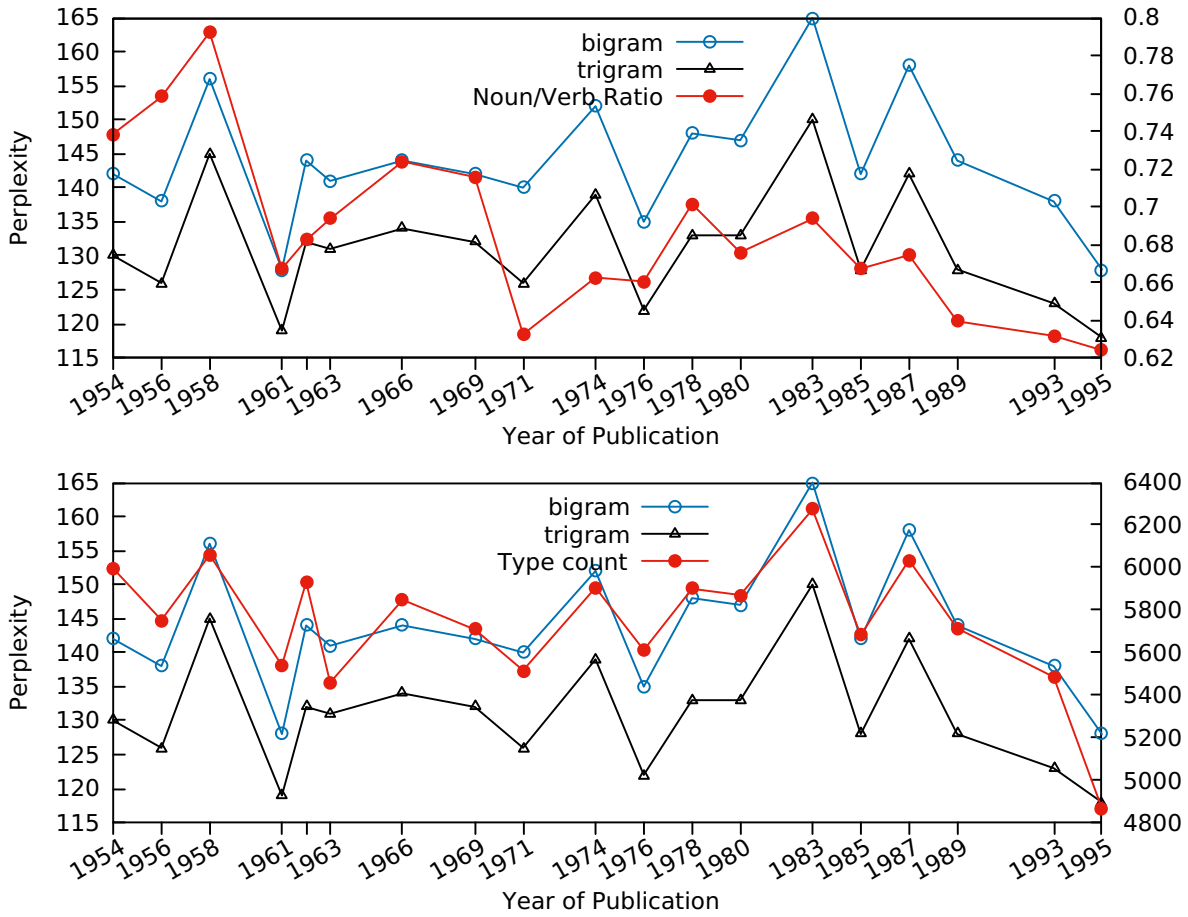

Figure 1: Perplexity values of the 19 novels, including direct discourse and Noun/Verb Ratio (above) resp. type count (below)

	Bigram&WittenBell	Trigram&WittenBell	Bigram&Kneser-Ney	Trigram&Kneser-Ney
Types	0.851	0.829	0.833	0.823
Noun/Verb Ratio	0.367	0.458	0.345	0.472

Table 1: Pearson correlation between the different measures across the 19 novels including direct discourse

One surprising data point is the 1961 novel "A Severed Head". Its perplexity values are similarly low to those of "Jackson's Dilemma". A possible explanation is that Murdoch used a simpler style intentionally as this novel is said to be more entertaining[12, Foreword]. As a consequence, it is reasonable to assume that the used language contains lots of dialog, short scenic descriptions, and less reflection of the character's emotional life.

As summarized in table 1, there is a high correlation between the number of types and the perplexity regarding both a bigram as well as a trigram model. Overall, the correlation values are only slightly influenced by the smoothing method and the order. Moreover, there is a medium correlation between the noun/verb ratio and the perplexity values.

Figure 2 shows the results where the sentences that are part of direct discourse have been skipped. With regard to the perplexity, it depicts what one would intuitively expect in a writer's career: a relatively simple language in the early works that evolves into a more complex language with a richer vocabulary.

The decreasing trend for the last three novels can be observed too, ending again in a perplexity value of the last novel which is significantly low. The overall higher count of types supports our assumption that the language excluding direct discourse is more diverse. There is again a high correlation between the number of types and the perplexity values as table 2 shows.

When skipping the direct discourse, the correlation between the noun/verb ratio and the perplexity vanishes. It is however notable that in the later novels, the overall trend of these measures is more similar than in the earlier ones.

Considering the works published in 1993 and 1989, it is striking that they both show a decreasing vocabulary and comparatively low perplexity. Although this connection is not statistically significant, it may nonetheless reveal very early signs of Alzheimer's disease.

5 Outlook

Beside the closed vocabulary approach chosen in this work, we could also define an universal background model of Murdoch's language and train the models with it as known words. For that purpose, we can use all the words that ap-

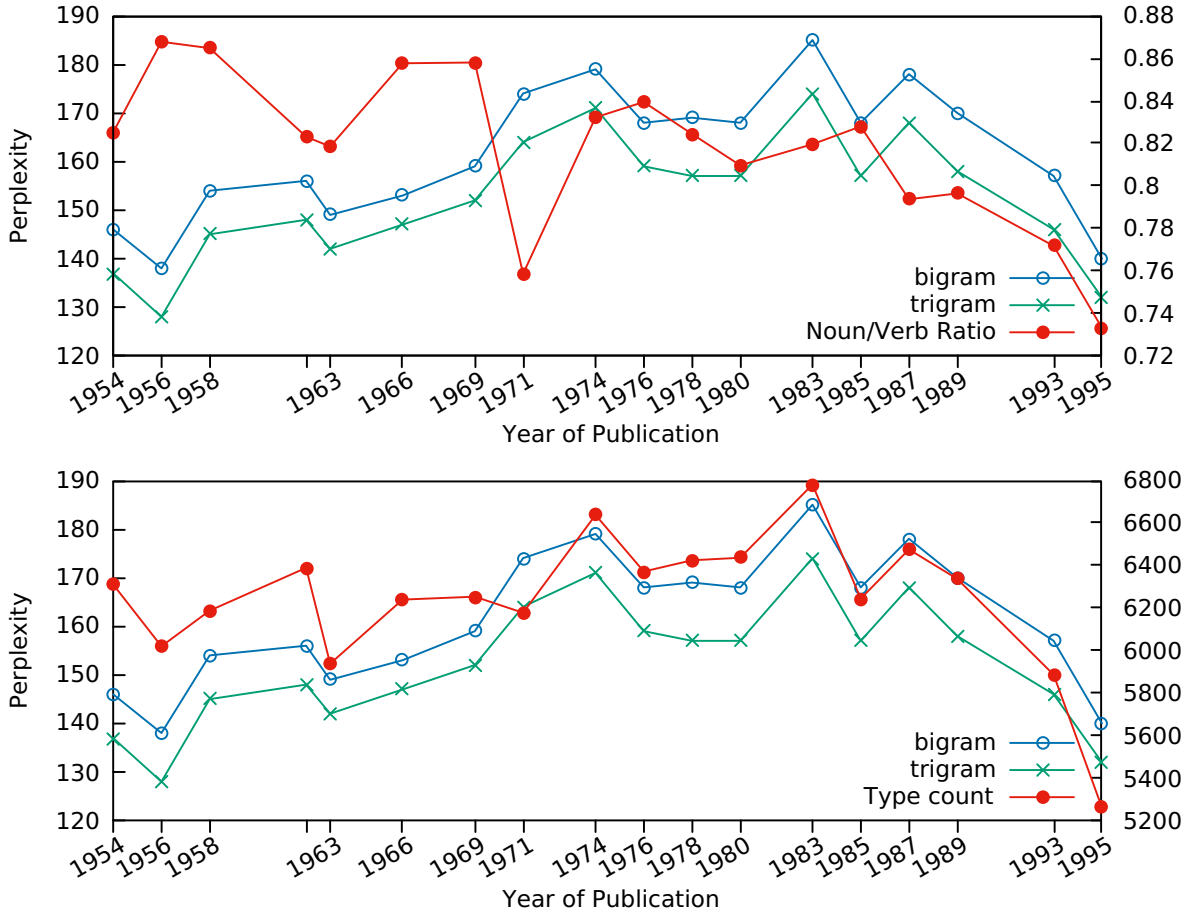

Figure 2: Perplexity values of the novels without direct discourse

	Bigram&WittenBell	Trigram&WittenBell	Bigram&Kneser-Ney	Trigram&Kneser-Ney
Types	0.747	0.745	0.748	0.752
Noun/Verb Ratio	-0.097	-0.072	-0.108	-0.077

Table 2: Pearson correlation between the different measures across the 18 novels without direct discourse

pear in $n\%$ of the novels. As a next step, we are going to investigate whether this approach delivers even better results.

Additionally, the investigations described in this work can be performed on other sources of language which are potentially influenced by Alzheimer's disease, too. A particularly considerable source is the Dementia Treebank[1] that contains transcriptions of picture descriptions by both Alzheimer's patients and a healthy control group.

In contrary to the texts by Murdoch, these descriptions come from a larger population of individuals with, presumably, less trained language skills. In addition to that, these picture descriptions are much shorter and were produced more spontaneously. We are going to determine whether these different texts show a similar pattern considering the perplexity values.

References

[1] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2015.

[2] P. Garrard, L. M. Maloney, J. R. Hodges, and K. Patterson, "The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author," *Brain*, vol. 128, no. 2, pp. 250–260, 2005.

[3] S. Pakhomov, D. Chacon, M. Wicklund, and J. Gundel, "Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing," *Behavior Research Methods*, vol. 43, no. 1, pp. 136–144, 2011.

[4] X. Le, I. Lancashire, G. Hirst, and R. Jokel, "Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists," *Literary and Linguistic Computing*, pp. 435–461, 2011.

[5] H. D. Spear, *Iris Murdoch*, ser. Modern Novelists. Palgrave Macmillan, 1995.

[6] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[7] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, ser. Practical Resources for the Mental Health Professionals Series. Prentice Hall, 2000.

[8] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J.

[1] http://talkbank.org/data/DementiaBank/

Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60.

[9] A. Stolcke, "Srilm-an extensible language modeling toolkit." in *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, 2002, pp. 901–904.

[10] T. C. Bell, J. G. Cleary, and I. H. Witten, *Text compression*. Prentice-Hall, Inc., 1990.

[11] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[12] I. Murdoch, *A Severed Head*, ser. Vintage classics. Random House, 2008.